DOCUMENT RESUME

ED 401 721 FL 024 173

AUTHOR Chalhoub-Deville, Micheline; Tarone, Elaine
TITLE Assessment Measures for Specific Contexts of Language Use.
PUB DATE 8 Oct 96
NOTE 35p.; Revised version of a paper presented at the Annual Meeting of the American Association for Applied Linguistics (18th, Chicago, IL, March 23-26, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Environment; Evaluation Criteria; *Language Proficiency; *Language Tests; Language Usage; Second Language Instruction; Second Language Learning; *Second Languages; *Teacher Made Tests; *Test Construction; *Testing; Test Interpretation; Test Use; Test Validity

ABSTRACT

A discussion of second language testing focuses on the need for collaboration among researchers in second language learning, teaching, and testing concerning development of context-appropriate language tests. It is argued that the nature of the proficiency construct in language is not constant, but that different linguistic, functional, and creative proficiency components are at work in different instructional and social contexts. Inadequacies of traditional and commercial tests for assessing contextualized language skills or determining instructional needs that are found frequently by teacher-researchers are examined. It is proposed that in both teaching and research, the validity of test score interpretation and use will be enhanced by use of tests constructed specifically for the instructional context in question, rather than generic, externally-produced proficiency measures. Broad criteria for construction of such measures are offered. Contains 100 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 721

Assessment Measures for Specific Contexts of Language Use

Micheline Chalhoub-Deville
Elaine Tarone

Center for Advanced Research on Language Acquisition
University of Minnesota
1313 Fifth Street, SE
Suite 111
Minneapolis, MN 55414
email: etarone@tc.umn.edu
Tel: (612) 627-1870
Fax: (612) 627-1875

October 8, 1996

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

chalhoub - Deville

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL024173

BEST COPY AVAILABLE

Abstract

The paper argues for more collaboration among the disparate areas of SLA research, L2 teaching and L2 testing in working toward common goals. All three areas are currently finding ways to consider the impact of social context upon L2 acquisition and performance.

Considerable work has been done in the area of SLA research exploring the impact of social interaction upon the development of interlanguage knowledge (e.g. Hatch 1978; Liu 1991; Tarone 1996). In L2 teaching, Tarone & Yule (1989), focusing upon needs assessment conducted by classroom second-language teachers, suggest highly local forms of assessment and research, which are descriptive of the language practices of specific individuals functioning in specific social contexts.

Finally, in L2 testing, the present authors contend that the nature of the proficiency construct is not constant but that different linguistic, functional, and creative proficiency components emerge when investigating the proficiency construct in different contexts. This paper will discuss the difficulties consistently pointed out by teacher-researchers regarding the inadequacy of traditional and ready-made assessment measures to assess learners' proficiency in acquiring such contextualized language skills, or to assist teachers in deciding what needs to be taught from one time to the next. This paper argues that teachers and researchers will be better served and the validity of their test score interpretation and use will be enhanced, if instead of employing generic imported proficiency assessment measures, they construct assessment measures according to the specific variables operating in their contexts of use (cf. Chalhoub-Deville, 1995a, 1995b; Turner & Upshur, 1995). Broad criteria for the construction of such measures are considered.

Micheline Chalhoub-Deville
Elaine Tarone
University of Minnesota

Introduction

A sign of a healthy field is its growth and expansion. Typically, as a field continues to grow, more and more specialization occurs. With increased specialization, however, researchers in the various areas of that field will need to exchange their findings to continue to foster growth in their respective areas as well as to pool their knowledge to advance the whole field.

Applied linguistics is clearly a healthy field. It has steadily expanded into several areas of specialization. As Ellis (1994) points out, however, this specialization within applied linguistics has made it challenging to keep abreast of developments in its various areas. Lack of communication among researchers in the various areas of applied linguistics is dangerous. It is likely to result in the continuous reinvention of the wheel, which, to say the least, does not enrich or advance the field. One goal of this paper is to show the benefits of more collaboration among areas of research in applied linguistics.

In the present paper we focus on three areas of inquiry in applied linguistics: research on second-language acquisition (SLA), second language (L2) pedagogy, and L2 testing. We begin with a brief argument for improved collaboration among researchers in SLA, L2 teaching and testing. Next, we outline some recent attempts to incorporate contextual effects into a theory of SLA. Then, we focus on the growing trend towards contextualized teaching, including the assessment of students' L2 needs in varied social contexts. We then explore standardized versus contextualized assessment, making a

---

* An earlier version of this paper was presented in March 1996 at the annual conference of the American Association for Applied Linguistics, Chicago, Illinois. A portion of this work was supported by the National Language Resource Center in CARLA (Center for Advanced Research on Language Acquisition) at the University of Minnesota.

case for the latter. Next, we discuss critical issues in contextual assessment, focusing mainly on tasks and rating criteria. Finally, we delve into the generalizability of such contextual assessment.

Research on SLA, L2 Teaching and Testing

Specialists in areas of applied linguistics such as SLA, L2 teaching, and L2 testing often seem unaware of one another's work, or at best, only superficially aware of work which is possibly related to their own. In general, research on L2 testing tends to be presented in separate conferences and published in separate journals from those read by SLA researchers and L2 teachers. Yet certain questions are of common interest to both SLA theorists and L2 testing theorists, such as: how can the knowledge of a second-language learner be modeled? In light of such obvious common interests, why don't SLA researchers and L2 testers cite each other more? One reason may be that work in both fields has burgeoned in recent years, so that keeping up with developments in each area alone is increasingly difficult. For example, Ellis' comprehensive 1994 overview of SLA research takes up more than 800 pages, with one chapter on applications of that research to L2 teaching, but only cursory mention of research in L2 testing. With regard to Bachman, a prominent language testing researcher with a well-known proficiency model whose work on the L2 construct extends back into the mid 70s, Ellis' only citation is: "There are alternative models of L2 proficiency (see Bachman 1990)..." (p. 24).

But the failure of communication is certainly not one-way. The language testing community has also failed to cite SLA research. To illustrate, Bachman's (1990) book, which provides an extensive documentation of the L2 testing research, does not refer to SLA researchers such as Ellis, Larsen-Freeman, Long, or Widdowson, or to critical SLA concepts such as variability theory. While the need to specialize is understandable, the tendency to ignore related research in the various areas cannot in the long run be healthy for the field.

Even attempts by one or another researcher in these respective fields to bridge the communication gap can be revealing of the magnitude of the gap itself. For example, Thomas (1994) attempts to document the way L2 proficiency (a central construct in L2 testing

research) is poorly operationalized in SLA research. Her "data come from a sample of literature published in relevant journals" (p. 307). The "relevant journals" are: *Applied Linguistics*, *Language Learning*, *Second Language Research*, and *Studies in Second Language Acquisition*, all central journals in SLA research. *Language Testing*, the lead source of publication for work in L2 testing, is not included. Since one of the major points of Thomas' paper is that SLA researchers' definitions of proficiency contain "inadequate or inappropriate information about proficiency that serves research poorly" (p. 307), it would behoove the SLA research community to examine how L2 testers define that concept in journals such as *Language Testing*. Similarly, L2 testers are urged to keep abreast of SLA research on the topic.

We emphasize that choosing to comment on these contributions is not intended to single out these authors, but simply to illustrate the phenomenon observed in applied linguistics in general: researchers in L2 teaching, testing, and SLA often seem to be working on related problems, but without much awareness of one another's work.

Of course there are good reasons for this phenomenon. The explosion of information in both areas has reached the point where it is all most of us can do to keep up with work in our own area; it may be beyond us to read research in related fields. But even if this is the case, we argue that collaboration between researchers from disparate fields can help to build bridges between those fields. It is our hope that this paper, a collaboration between an SLA specialist and an L2 testing specialist, both of whom are involved in L2 teacher preparation, will serve as an impetus to motivate closer collaboration and more communication among the different areas of research in applied linguistics to the benefit of those areas and the whole field.

To begin, we will point to the work of a group of SLA researchers who take the position that SLA theory should include a description and explanation of the impact of social context, including social interaction, upon the development of the L2 learner's interlanguage.

Social Context and SLA Theory

Hatch (1978), Selinker and Douglas (1985) Preston (1989), Gass (in press) and others have taken the position that the study of SLA should include research which examines the impact of social interaction upon the internal development of an interlanguage grammar. According to Tarone (1996a,b), one of the central questions of such research should be: CAN internal cognitive processes of SLA be affected by social interaction and social context, and if so, HOW? Tarone (1983, 1988) and other variationists (e.g. Dickerson 1975; Ellis 1985, 1987; Young 1991) focus their research upon L2 learner performance in a variety of social contexts, believing interlanguage variation across those contexts to be importantly related to change in learners' IL knowledge over time. Variation, from this perspective, can be a source of information about the way in which interaction in different social contexts can influence both interlanguage use AND, potentially, overall interlanguage development. Tarone (1983, 1988, 1990) has argued that it is important for any SLA theory to describe and explain why it is that interlanguage performance varies systematically from one social context to another, and to relate this variation in performance to the development of the interlanguage system. Research evidence from studies such as Liu (1991) (also described in Tarone and Liu 1995) supports this view. Liu's longitudinal case study of a child learner of English L2 showed that the learner's progress through several stages of acquisition of English questions was affected substantially by interactional context; indeed, Liu argues that interactional forces interacted with cognitive forces so strongly as to alter supposedly universal sequences of development. The viewpoint that a theory of SLA should include some account of the effect of social context and social interaction upon interlanguage development is attracting considerable support from such researchers as Ellis (in preparation), Gass (in preparation), Mitchell, Hooper, and Miles (in preparation), Young (in preparation), Olshtain (in preparation), and Tarone and Beebe (in preparation).

How is this trend in SLA research paralleled by current trends in L2 teaching?

L2 Teaching

Our basic assumption in this paper is that the entire L2 teaching enterprise must start and end with specific L2 learners who must function in the L2 in specific local social situations. Highly pragmatic, English as a second language (ESL) teachers attempt to teach THEIR students the English THEY need to know. They adapt generic ESL textbooks to meet the students' local needs, as nearly as those needs can be established.

How do ESL teachers establish what aspects of English their students need to learn? Tarone and Yule (1989) point out that classroom second-language teachers are constantly involved in a highly local, ongoing process of needs assessment: establishing what their learners know of the L2 (SLA research) and what they need to know (e.g., English for specific purposes (ESP) research). This assessment of student needs by teachers is always approximate, limited by the teachers' time and energy. It is local, everyday in-class assessment by teachers for teaching purposes.

Given more time, as in a graduate-level teacher training program such as the M.A. program in English as a Second Language (ESL) at the University of Minnesota, such teachers, retaining their pragmatic attitude, produce highly local forms of assessment and research, which are descriptive of the language practices of specific individuals functioning in specific social contexts. This paper will present examples of these studies, many of them carried out by M.A. level ESL teachers, which illustrate the extreme variation in the registers and language skills needed in such different social contexts as the doctor-patient interview, the welfare office, the telephone, the basketball court and the chemistry lab.

The existence of this sociolinguistic variation, documented in these sorts of "ESP" studies, and completely consistent with the contextual accounts of SLA described earlier, will lead to the following claim: the nature of the language proficiency construct is not constant; different linguistic, functional, and creative proficiency components emerge when we investigate the proficiency construct in different contexts. We can think of no social situation in which one draws equally on ALL aspects of one's proficiency in a language.

While that proficiency may be there in theory, different aspects are needed and so are used differentially in specific social situations. Further, no speaker ever participates in every possible social context in any culture -- thus every speaker develops some aspects of "English language proficiency" more than others.

The following studies, then, show that different social situations call for the L2 learner to use some aspects of proficiency MORE than others. We argue that because of this need s/he will develop some areas of proficiency more than others.

M.A. students in ESL at the University of Minnesota are required to write qualifying papers at the end of their coursework. At the time they are asked to do this work, all are simultaneously teaching international students at the University and deeply involved in practical classroom issues. All want to write USEFUL papers. They usually do needs assessments, often focusing on ESP -- describing the way English is used in different social contexts where their students need to function. As a result, we have a growing set of descriptions of the English language practices of specific individuals functioning in a variety of specific social contexts. Based on these descriptions, classroom tasks are developed to train ESL learners to perform in authentic communication. We will now describe some of those papers.

Several (Levine (1981), Ranney (1992), Mori (1991), describe the doctor-patient office interview, both in terms of what NNSs need to know and what they in fact know. What emerges from these studies is that NNSs of English need a variety of oral skills (sociolinguistic skills, negotiation skills, and vocabulary). First, non-native speakers (NNSs) need to share the same script (set of sociocultural expectations) as the doctor as to what the goal of the interaction is; the doctor typically thinks the goal is to reach a diagnosis, or understanding of the nature of the problem, but the NNS often thinks the goal is not to obtain a diagnosis but rather a prescription: some concrete medication to take out of the office. Another part of the script that must be shared involves what sort of evidence the doctor will be trying to collect during the course of the interview (direct measurements of temperature and blood pressure

as well as statements from the patient, vs. direct visual, tactile and olfactory clues). In addition to the above, the patient will need to know that the doctor is under substantial time pressure to get in and out of the office as quickly as possible; in spite of this the patient must use negotiation skills and assertiveness in making clarification requests and confirmation checks. Finally, the patient will need health-related vocabulary to explain symptoms, and receptive understanding of English directives and recommendations. (One patient (Levine 1981) didn't know the word "dizzy" and so couldn't explain that his heart medication had that side effect.) Aspects of English language proficiency NOT usually needed in this social context include reading/writing skills; formal oral presentation skills; or in grammar, and use of the future tense, among others. Recommended classroom instructional tasks include tasks focusing on use of vocabulary to describe symptoms, oral negotiation tasks using clarification requests and confirmation checks under time pressure, and explicit comparison of scripts for doctor-patient interviews.

We can turn now to another social setting which has been the object of study in our program. "Survival English" textbooks cover many situations in which new arrivals need language support: the post office, the bus, the store. One situation which is commonly encountered by recent immigrants, but which never turns up in survival English curricula, involves the social services office. A recent study (Kuehn 1994) taped and described NS and NNS clients as they went through a welfare office intake interview in applying for social services to which they were entitled in rural Minnesota. This teacher researcher had taught in rural Minnesota and had always had a number of recent immigrants who were legally qualified for social services but who had a very difficult time with intake interviews. Fortunately this teacher had also worked as an intake interviewer in thewelfare office and so was able to get permission to tape, transcribe and analyze 2 interviews. She was able to identify a highly ritualized prescribed script used in the social services financial intake interview, in which there were 3 major transactions, all areas in which the NNS had language-related difficulties. The greatest difficulties were in understanding the

structure of the script, and the jargon used by the interviewer, and consequently in responding to confirmation requests and understanding directives. This study recommended several tasks which might be used in the classroom to better prepare students to deal with this social context.

Another study which we would like to describe is by Rimarcik (1996), who discovered that her students were having tremendous difficulties in listening comprehension in a previously unidentified social context: listening and responding to automated voice response systems (AVRSs) on the telephone. AVRSs are those computerized systems which answer the phone, list options for choice and ask you to press 1 if you want the doctor, 2 if you're really sick, and 3 if you're dead. This context is not covered in ANY commercial ESL textbook (or, we assume, assessment instrument), and yet it is ubiquitous these days for anyone who needs to use the phone. Rimarcik taped 12 messages, transcribed them, analyzed their logical and linguistic structure, and then used them to design instructional tasks for her learners. She found that these messages imposed substantial logical and memory burdens on her students. Interestingly, the AVRS which was longest, most complex linguistically and most difficult to process cognitively was the one which was supposedly aimed at immigrants: the INS message system. In listening, Rimarcik found that her students needed to understand the use of several variants of the conditional:

If you wish/want/are/would like X, press Y.

For X, press Y.

To X, press Y.

If you have N, press Y.

They also needed to know terms like "pound key" and "star key".

In addition to these studies, there are studies of interactions in university physics labs (Jacobson 1992), of lecture note-taking in business classes (Schmidt 1981), of politeness strategies in written business letters (Maier 1992), even of English language use, including "trash talk", on the basketball court (Trites 1996).

Our point here, quite simply, is that each of these different social contexts requires a different configuration of various

components of language proficiency. ESL teachers want to be able to identify these configurations in order to better tailor their teaching to the needs of their students.

With such a highly-contextualized approach to teaching, and given the need to align teaching and assessment, we would like to ask, what approach to assessment is most suitable? Teachers favor a contextualized assessment approach for both in-class and for outside-the-classroom purposes. What issues need to be considered in the development of contextualized assessment?

Assessment Issues

Effects of Standardized Testing on Curriculum

How can we assess the proficiency of these adult ESL students who need to use the language in specific, real-life sociolinguistic contexts? The facile solution is to rely on a standardized, or off-the-shelf test. Although such tests are attractive because they are readily available, the literature cautions us against using such tests for two primary reasons. First, standardized tests usually focus on generic proficiency that is supposedly transferable to all contexts. As is repeatedly argued in this paper, different contexts have complex and dynamic qualities and standardized tests do not recognize or necessarily accommodate these different contexts. This issue is treated in more detail later in the paper.

Second, these standardized tests have grave effects on the curriculum. Madaus (1988), Mehrens and Kaminski (1989), and Smith (1991) examine the impact of standardized testing on the curriculum and report that often the curriculum is being geared to the test rather than the test being geared to the curriculum. The tests, as such, are defining the objectives of the teaching/ learning situation and forcing classroom teachers to subjugate their lesson plans to test preparation. Smith (1991) maintains that standardized tests "substantially reduce the time available for instruction, narrow curricular offerings and modes of instruction, and potentially reduce the capacities of teachers to teach content and to use methods and materials that are incompatible with standardized testing formats" (p. 8). Teachers are typically anxious to prepare students for the

tests because test results are often used as indicators of the quality of their teaching.

Thus, many authors lament the unfortunate fact that standardized tests seem to work against the local enterprise, to sabotage efforts to meet students needs. Such tests tend to discourage innovative or creative approaches to teaching. In addition, this type of assessment does not usually allow to focus on testing the language students are expected to use. What is needed is to get assessment in line with the contextualized approach to teaching, described earlier in this paper. Components salient to a particular context need to be the focus of assessment.

Proficiency

A primary step in assessment is defining the construct being measured. A survey of the testing literature shows that no single definition of L2 proficiency is accepted. The different ideologies and purposes have led to the development of models with varied representations of the proficiency construct. The models vary profoundly in the breadth of their components, ranging from the single component, e.g., Oller's (1976) to the multiple components, such as Bachman's (1990). For a review of some of these models see Chalhoub-Deville (forthcoming b) and Skehan (1987). Briefly, the literature indicates researchers' preference for multi-componential models. These multi- componential models, however, afford researchers diverse representations of the nature of the L2 proficiency construct.

The lack of consensus in portraying the nature of L2 proficiency has prompted researchers such as Lantolf and Frawley (1985, 1988, 1992) to argue that valid assessment cannot be achieved without a commonly accepted model of proficiency. Spolsky (1992, in North 1993) argues that the search for this one model resembles that of looking for the unattainable "holy grail." Researchers such as Chalhoub-Deville (forthcoming b), Henning and Cascallar (1992), and Spolsky contend that no single model can serve the diverse purposes of assessment. Any given model may be suitable for certain contexts, but not for others. To

illustrate we focus on the communicative language ability (CLA) model by Bachman (1990) and Bachman and Palmer (forthcoming).

The CLA model has been claimed an advance in the representation of the proficiency construct. The CLA model is based on empirical research, mainly the Bachman and Palmer (1982) study, and the theoretical contributions of Hymes (1972, 1973), Munby (1978), Canale and Swain (1980), Canale (1983), and Savignon (1983). The CLA model is too complex to be summarized in its entirety here, but for now we can describe it as modeling the learner's knowledge as consisting of three broad components: schemata of knowledge about the world, language knowledge, and affective schemata. Language knowledge is further subdivided five components: grammatical, textual, lexical, functional and sociolinguistic knowledge. The model provides still further details of all the various aspects of each of these components of language knowledge (see Table 1).

insert Table 1 here

Such detail in the representation of the proficiency construct is quite informative. Researchers such as McNamara (1990), Skehan (1991), and Chalhoub-Deville (forthcoming b), however, contend that the model is too inclusive, which makes it hard to implement in its entirety. The dilemma that arises here is the need, on the one hand, for complete models that provide a comprehensive representation of the construct, and the challenge, on the other, of implementing such models. Such a dilemma can be resolved by distinguishing between theoretical models that emphasize completeness and operational models that underscore parsimony.

In general, theoretical models purport to define proficiency at a general level across contexts. Operational models are usually based on theoretical models, but are not all-inclusive. Operational models reinterpret theoretical models to focus on the specific needs or variables operating in a given context of use. For test development purposes, it is more appropriate to convert theoretical models into operational models that portray the construct at a contextual level.

Table 1: Bachman & Palmer's (forthcoming)
Theoretical Model of Communicative Language Ability

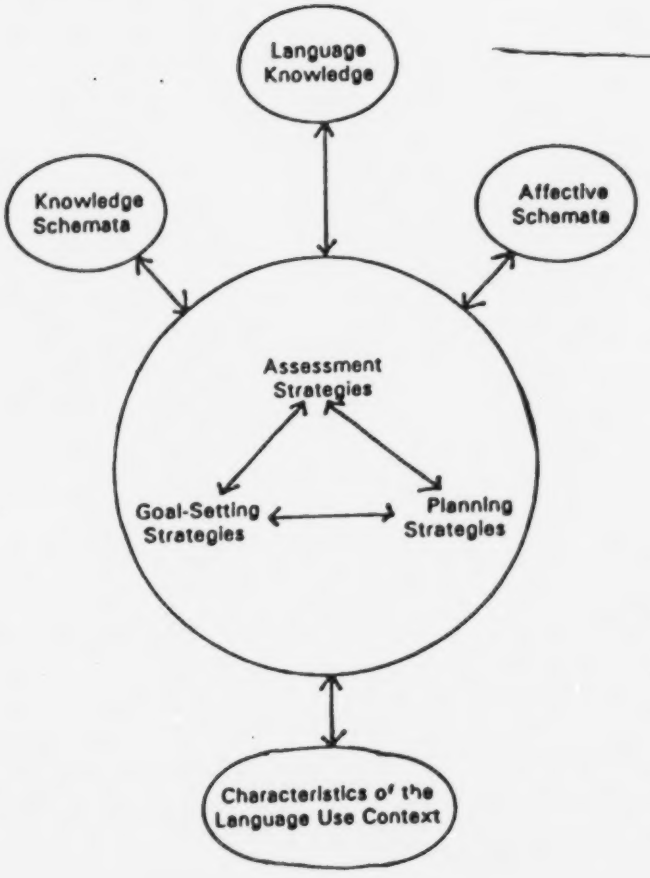

Figure 3.1

A Model of Language Use

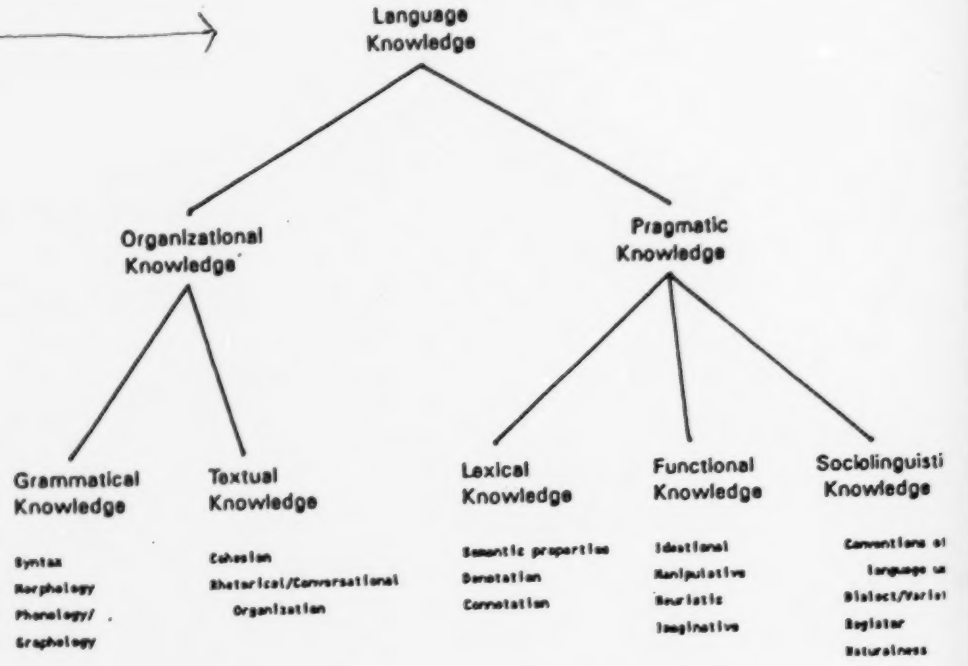

Figure 3.2

Areas of Language Knowledge

BEST COPY AVAILABLE

Below we provide an example that illustrates how to recast a theoretical model into an operational one that accommodates the particular context.

Contextualized Assessment

In the first section of this paper, we described several social contexts in which ESL learners need to perform effectively and use their proficiency. We pointed out that in those situations, the learners did not need to draw equally on their proficiency in all aspects of the English language; rather each situation seemed to call for differential use of different registers, skills, and grammatical structures of English.

An operational model of proficiency which might apply to the situation of the doctor-patient interview might focus upon these components of the Bachman and Palmer model:

Knowledge Schemata: the learner needs to know the script for American doctor-patient interviews

Language Knowledge:

Pragmatic Knowledge

Sociolinguistic Knowledge:

Conventions of Language Use: turn-taking

Register: medical

Lexical Knowledge: vocabulary to describe symptoms

Functional Knowledge

Manipulative: ability to understand medical directives

Metacognitive Strategies: ability to undertake oral negotiation using clarification requests and confirmation checks under time pressure.

Thus, in considering the skills the learner needs to use in a doctor-patient interview, an operational model of proficiency will specify only some of all of the components in the Bachman and Palmer theoretical model of proficiency, as illustrated above. It seems clear that the patient's grammatical accuracy in this situation will be less important than the patient's pragmatic, metacognitive, and world knowledge abilities.

In addition to identifying the contextually appropriate language components, we need to carefully consider both the tasks

Table 2: An Operational Model of Communicative Language Ability Most Needed for a Doctor-Patient Interview

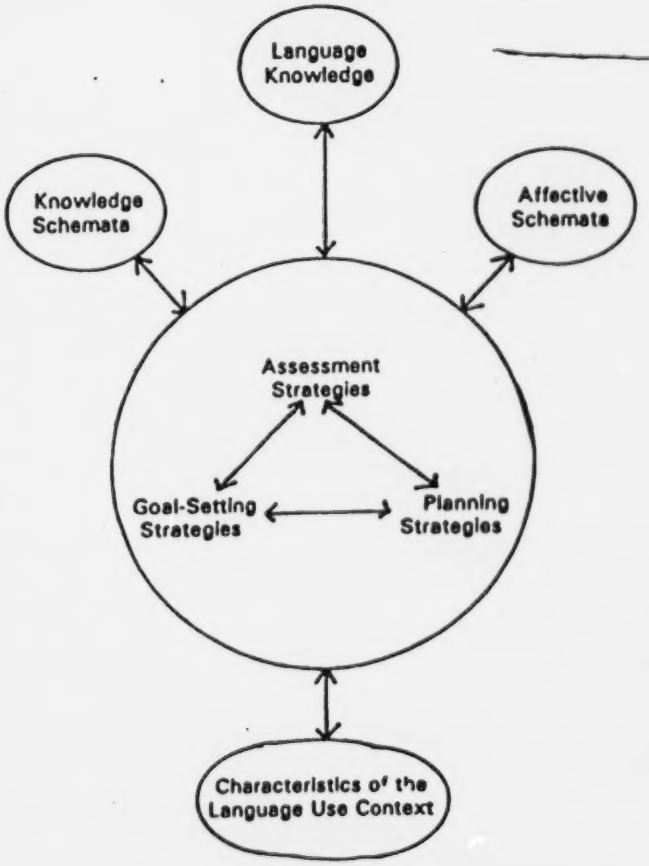

Figure 3.1

A Model of Language Use

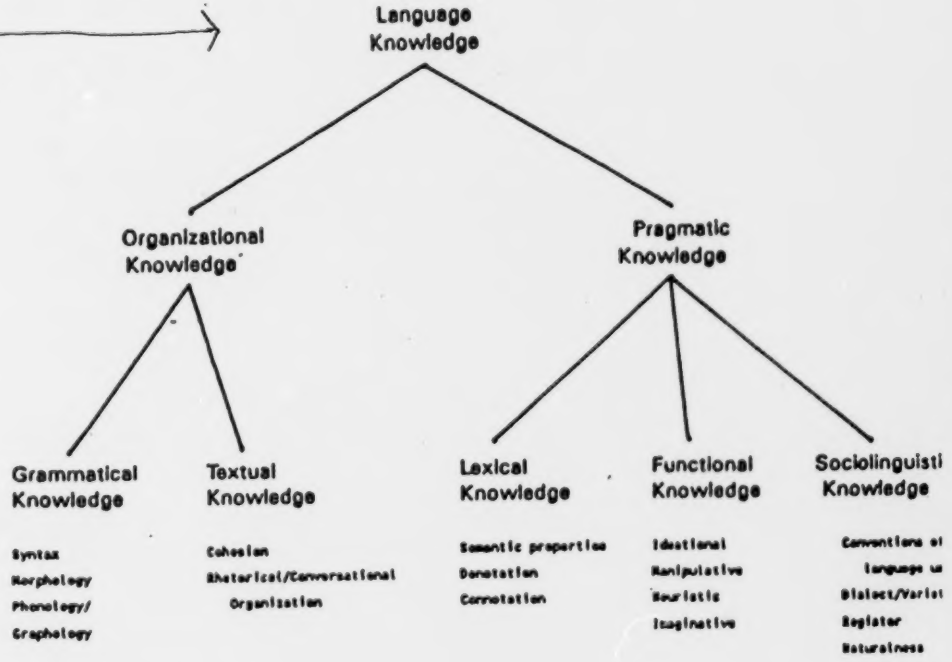

Figure 3.2

Areas of Language Knowledge

BEST COPY AVAILABLE

that tap those selected components and the criteria that focus on those components. It is imprudent to contextualize the components, but then to randomly select the tasks and the corresponding rating criteria. Both the tasks and the criteria need to be selected in accordance with the specifics of a given context.

Elicitation Tasks

Both SLA and L2 testing research document the variable performance of students on different tasks (Brown and Yule 1983). SLA studies demonstrate that different tasks engender variable output in lexicon (Pavesi 1987), phonology (Beebe 1980, Dickerson 1975, Schmidt 1977, Sato 1985), morphology (Larsen-Freeman 1975), and syntax (Schmidt 1981). (For a comprehensive listing of this SLA research, see Tarone 1988.)

Similarly, L2 testing research has shown that student performance is not constant, but varies across tasks. Studies such as those by Bachman and Palmer (1981), Clifford (1981), Henning (1983), Shohamy (1983), Shohamy, Reves, and Bejarano (1986), Wolf (1993), and Chalhoub-Deville (1995a) document what is sometimes called the "method effect": the way learners' varied performance leads to diverse scores on different tasks.

Such variability can be attributed largely to the different demands that the task places on the linguistic and cognitive processes of the learners, thus, influencing their performance. For example, with respect to the interview task, the interviewer is present to interact with the learners and to direct their efforts in constructing speech. In the read-aloud, learners are provided with a text that obviously constrains their language production and does not allow for interaction with another speaker or for immediate feedback. (See Brown and Yule 1983, and Tarone and Yule 1989, for discussions on the impact of interaction with another speaker on task performance.)

The documented SLA and language testing variability has led Tarone (1983), Ellis (1985, 1990, 1994), Bachman (1990), and Larsen-Freeman and Long (1991) to prompt researchers to sample varied tasks to elicit a range of learners' L2 proficiency. Indeed sampling a variety of tasks affords researchers a richer picture of

learners' proficiency. Considering, however, that research, e.g., Chalhoub-Deville (1995b) provides evidence that different components underlie diverse tasks, it is imprudent to randomly sample a wide range of tasks. Researchers are prompted to consider the tasks that are likely to tap their intended proficiency components. The Chalhoub-Deville (1995b) study examined students' performance on three tasks: an interview, a narration, and a read-aloud. Analyses indicated that grammar-pronunciation and appropriate vocabulary usage were the salient components in the interview samples, creativity in presenting information and grammar-pronunciation were prominent for the narration samples, and finally confidence and pronunciation emerged in the read-aloud samples.

To sum up: the observed nature of the L2 proficiency construct is not constant across tasks. Different tasks are likely to capture different aspects of learners' proficiency. Researchers are encouraged not only to sample a variety of tasks to capture a richer picture of learners' performance, but also to sample tasks that ensure the prominence of the required proficiency components.

Rating Criteria

We have argued for closely matching the proficiency components with the intended context. We have also contended for selecting tasks that tap the salient components in that particular context. In this section we make a case for the contextual development of the evaluation criteria. It is contradictory and self-defeating to contextualize the proficiency components and the tasks and then to use some generic criteria for evaluation.

For two decades applied linguists have been occupied with learners' errors and with how those errors are perceived by various NSs and NNSs, i.e., error evaluation (Ellis 1994). Error evaluation studies and reviews such as those by Albrechtsen, Henriksen, and Faerch (1980), Chastain (1980), Eisenstein (1983), Davies (1983), Guntermann (1978), Ludwig (1982), Magnan (1982), Piazza (1980), and Politzer (1978) have tended, as Brindley (1991) writes, to investigate "the effects of particular discourse, phonological, syntactic or lexical features

on comprehensibility and/or irritation, rather than relating them to perceptions of proficiency" (p. 156). Although concern with errors is informative, it is not sufficient. Research is needed that focuses on the overall perception of learners' proficiency.

Proficiency research shows that rater groups with diverse professional training and background experiences differ in their expectations and evaluations of students' proficiency. Research (Ervin 1977, Fayer and Krasinski 1980, Galloway 1980, Barnwell 1989, McNamara 1990, Hadden 1991, Schairer 1992, Brown 1995 Chalhoub-Deville 1995a, 1995b, Elder 1996) documents differences between NSs and NNS, between teachers and non- teachers, and between NSs whose place of residence is the L1 community vs. those NSs in the L2 community. These differences are not only in terms of the scores awarded, but also with regard to the proficiency components raters elect to focus on when observing learners' performance.

In short, the above studies provide a strong evidence against the generic conceptualization of the NS. The question that consequently arises is what criteria, or more appropriately whose criteria should be used in evaluating learners' performance? We contend that it depends on who the end-user of the results are, i.e., the context in which these results are used. "Accurate interpretation and use of test scores necessitates the inclusion of criteria that correspond to the perceptions of the end-users" (Chalhoub-Deville forthcoming a). To explicate this point about the rating criteria and the end- user, we focus on the evaluation of ESL learners in varied contexts of use.

First we would like to consider the assessment of ESL students for classroom use. In such a context, students are being tested in order to inform further instruction and ESL teachers are typically the end-users of the test results. In other words, teachers use the test results to design/adjust the subsequent lesson plans. Criteria, therefore, should be congruent with the language components deemed important by those ESL teachers. Alderson and Clapham (1995), Upshur and Turner (1995), and Turner and Upshur (1995) similarly argue that in order to obtain appropriate and meaningful

assessment of learners' performance, classroom assessment criteria should include teachers' views and beliefs. If rating criteria do not reflect what teachers deem important, there are no guarantees that the results obtained will be interpreted and used appropriately. In short, for effective classroom teaching, assessment in general and the rating criteria used specifically need to consider the L2 components teachers deem meaningful and appropriate.

In this second example, we consider the rating criteria when the test is intended to measure learners' ability to perform in academic settings, e.g., pursuing studies where ESL is the medium of instruction. In such a context, rating criteria should reflect the perceptions held by those teachers with whom students are expected to interact in their academic work. It is the perceptions of those non-ESL teachers that are contextually more pertinent. The ESL teacher cannot be expected to be familiar with how the language is used to discuss the various academic subject matters. Those non-ESL teachers are better equipped to determine learners' ability to use ESL for academic purposes. As mentioned before, if the perceptions of the appropriate group of teachers are not included, the interpretation and use of proficiency ratings are jeopardized.

As for ESL assessment for professional certification, evaluation criteria should take account of the views held by representatives of that professional community and not necessarily of ESL teachers. Brown (1995) forwards a similar argument. She states that given the differences in the rating behaviour between the teachers and tour guide professionals, and given the context of professional certification, criteria reflecting the perceptions of the tour guide representatives are likely to be more appropriate than those of the L2 teachers. It is those professionals and not necessarily the ESL teachers who have the pertinent knowledge and intuition of the proficiency deemed appropriate in the targeted professional setting(s).

To summarize what we have shown thus far: a discrepancy between the intended context of assessment and the rater group employed to derive the rating criteria threatens the meaningfulness and usefulness of the rating results. We contend that rating criteria

should be based on the perceptions of the rating group(s) most consistent with the particular context(s).

Generalizability

In advocating contextual assessment, where the selection of proficiency components and the corresponding development of tasks and evaluation criteria are context dependent, the generalizability of the ensuing results becomes critical to address. Likewise, Messick (1989) writes: "Because of numerous factors contributing to interactions and systematic variability in behavior and performance, generalizability of the construct meaning of test scores across various contexts cannot be taken for granted" (Messick, 1989, p. 56). The generalizability issue is indeed of paramount importance. We do not propose to provide answers that settle this complex issue. We do, however, forward a couple of perspectives for the reader to consider.

In addressing the generalizability issue, we ask the reader to carefully consider the other side of the argument, i.e., if generic assessment is used, to what contexts do the scores obtained generalize? As Chalhoub-Deville (forthcoming b) maintains, "while recognizing that such context-specific assessment frameworks may lack generalizability, the mindful practitioner would also recognize that it is imprudent to promote generalizability at the expense of validity." We contend that with contextualized assessment we have a more accurate representation of learners' proficiency in that specific context.

Another critical issue to consider in this discussion of generalizability is the relationship between theoretical and operational models. By linking operational models to theoretical models, the researcher can judge how the proficiency components of that delimited and context-dependent operational model fit into the more generic theoretical model. Such an approach enables deliberation and discussion about the meaning of the proficiency construct across contexts.

This approach is certainly congruent with the approach taken by many SLA researchers; as Ellis (1994) says:

The object of our enquiry--second language (L2)

> acquisition--is best seen as a complex, multi-faceted phenomenon--more like a many-sided prism than a neat picture with clearly identifiable objects. The images that the prism presents vary in accordance with the angle from which it is viewed and the light directed at it, with the result that, while they are in some way interrelated, they also afford different perspectives of the same entity. (p. 667)

In the domain of L2 testing, we might put it this way: proficiency does not manifest itself as one unchanging set of components. Instead, proficiency denotes subsets of components depending on the variables operating in a given context. The subsets provide snap shots of learners' proficiency. These snap shots are interrelated and together they provide a rich and multi-faceted picture of the proficiency construct.

Conclusion

The paper underscores the need for applied linguists in the various areas of research to collaborate to foster growth in the field.

We have focused on the need to align the related areas of L2 teaching, SLA research, and L2 testing and have made a case for both contextual teaching and testing. In the area of SLA research, we have pointed to work which suggests that SLA theory should account for the impact which social context and social interaction have upon the learner's development of an interlanguage. In the area of teaching, we have seen that careful analysis of the target situations can lead to identification of language components and tasks that learners need to learn. Those components and relevant tasks can be used in the classroom to teach students the skills they need.

With regard to assessment, we have argued that the context with its particular purpose, language, examinee, task, rater, etc., causes certain components of the L2 construct to be relevant and others irrelevant. As a result, we have advocated, as more appropriate for testing, the use of operational models that include only the contextually salient components. Furthermore, we have made a case both for the careful selection of tasks that tap the

appropriate language components and for developing evaluation criteria that reflect the perceptions of the potential users.

In our discussion, we have tried to address the danger that by localizing we may lose generalizability. By developing an endless stream of local assessment instruments, we may lose the ability to determine whether a learner who has become proficient in communicating in one situation has also become proficient in communicating in another. We have suggested that when each operational model is tied to a comprehensive theoretical model, then we can gain some ability to generalize beyond the specific local situation, to other situations in which the same or similar constellations of proficiency are called for. Nevertheless, research is needed, as called for also by Baron (1991), Dunbar, Koretz, and Hoover (1991), and Linn and Burton (1994), to investigate the number and types of tasks needed to provide an adequate representation of learners' proficiency.

References

Albrechtsen, D., Henriksen, B., & Faerch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning, 30*(2), 365-396.

Alderson, C. & Clapham, C. (1995). Assessing student performance in the ESL classroom. *TESOL Quarterly, 29*, 184-187.

Bachman, L. F. & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp.149-165). Washington, DC: Teachers of English to Speakers of Other Languages.

Bachman, L. and A. Palmer (forthcoming) *Language Testing in Practice.* Oxford: Oxford University Press.

Baird, K. (1991, August 9). Company teaches test-taking. *Columbus Dispatch*, p. 5C.

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*, 152-163.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305-318.

Beebe, L. (1980) Sociolinguistic variation and style-shifting in interlanguage. *Language Learning* 30,2:433-447.

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: Regional Language Center.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*, 1-15.

Brown, G. and G. Yule (1983) *Teaching the Spoken Language.* Cambridge: Cambridge University Press.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller, Jr. (Ed.), *Issues in Language Testing Research* (pp. 333-342). Rowley, MA: Newbury House Publishers.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Chalhoub-Deville, M. (1995a). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-33.

Chalhoub-Deville, M. (1995b). A contextualized approach to describing oral language proficiency. *Language Learning, 45*, 251-281.

Chalhoub-Deville, M. (forthcominga). Theoretical models, assessment frameworks, and test construction. *Language Testing.*

Chalhoub-Deville, M. (forthcomingb). *Test Interpretation, Test Use, and Pedagogical Implications.* Australian Review of Applied Linguistics.

Chastain, K. (1980). Native speaker reaction to instructor-identified student second-language errors. *Modern Language Journal, 64*, 210-215.

Clifford, R. T. (1981). Convergent and discriminant validation of integrated and unitary language skills: The need for a research model. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp.149-165). Washington, DC: Teachers of English to Speakers of Other Languages.

Davies, A. (1983). Error evaluation: The importance of viewpoint. *English Language Teaching Journal, 37*(4), 304-311.

Dickerson, L. (1975) The learner's interlanguage as a system of variable rules. *TESOL Quarterly* 9:401-407.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Eisenstein, M. (1983). Native reactions to non-native speech: A review of empirical research. *Studies in Second Language Acquisition*, 5(2), 160-176.

Elder C. (March, 1996). Performance testing for the professions: Language proficiency or strategic competence. Paper presented at the annual meeting of the American Association for Applied Linguistics, Chicago, IL.

Ellis, R. (1985) *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (1987) *Second Language Acquisition in Context.* London: Prentice Hall International.

Ellis, R. (1990). *Instructed second language acquisition.* Cambridge, MA: Basil Blackwell.

Ellis, R. (1994) *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, R. (in preparation) Item Learning vs. System Learning: Explaining Free Variation. Presentation for AAAL 1997 Colloquium: Incorporating Sociolinguistic Perspectives into SLA Theory.

Ervin, G. L. (1977). *The study of the use and acceptability of target language communication strategies employed by American students of Russian.* Unpublished doctoral dissertation, The Ohio State University.

Fayer, J. M. & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-326.

Gass, S. (in press) *Input, Interaction and the Development of Second Languages.* Lawrence Erlbaum Associates.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*, 428-433.

Gass, S. (in preparation) An Input/Interaction Perspective on Acquisition. Presentation for AAAL 1997 Colloquium: Incorporating Sociolinguistic Perspectives into SLA Theory.

Guntermann, G. (1978). A study of the frequency and communicative effects of errors in Spanish. *Modern Language Journal, 62*, 251-253.

Hadden, B. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning, 41*(1), 1-24.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Hatch, E. (1978) Discourse analysis and second language acquisition, in E. Hatch (Ed.) *Second Language Acquisition: A Book of Readings.* (pp. 402-435) Rowley, Mass: Newbury House.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning, 33*(3), 315-332.

Henning, G. and Cascallar, E. 1992: *A preliminary study of the nature of communicative competence.* Princeton, NJ: Educational Testing Service.

Hymes, D. (1972) On communicative compentence, in J. Pride and J. Holmes (Eds.) *Sociolinguistics.* Harmondsworth: Penguin.

Hymes, D. H. (1973). Toward linguistic competence. *Texas Working Papers in Sociolinguistics, Working Paper No. 16.* Austin, Texas: Center for Intercultural Studies in Communication, and Department of Anthropology, University of Texas.

Jacobson, W. (1986) An assessment of the communication needs of non-native speakers of English in an undergraduate physics lab. *English for Specific Purposes* 5,2:173-188.

Kenyon, D. & Stansfield, C. (1992). Examining the validity of a scale used in a performance assessment from many angles using the many-faceted Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kuehn, K. (1994) Form-structured discourse: A script of a welfare office intake interview for ESL learners. MA Qualifying Paper, University of Minnesota.

Lantolf, J. P. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal, 69*, 337-345.

Lantolf, J. P. & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*, 181-195.

Lantolf, J. P. & Frawley, W. (1992). Rejecting the OPI--again: A response to Hagen. *ADFL Bulletin, 23* (2), 34-37.

Larsen-Freeman, D. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly, 9*, 409-430.

Larsen-Freeman, D. and M. Long (1991) *An Introduction to Second Language Acquisition Research*. London: Longman.

Levine, L. (1981) Influence of social role on communication in a cross-cultural interview. MA Qualifying Paper, University of Minnesota.

Linn, R. L. & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), pp. 5-15.

Liu, G. (1991). Interaction and second language acquisition: A case study of a Chinese child's acquisition of English as a Second Language. Unpublished Ph.D. Thesis, La Trobe University.

Ludwig, J. (1982). Native-speaker judgments of second language learners' efforts at communication: A review. *Modern Language Journal, 66*, 274-283.

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76.

McNamara, T. F. & Lumley, T. (1993). The effects of interlocutor and assessment mode variables in offshore assessment of speaking skills in occupational settings. Paper presented at the annual meeting of the Language Testing Research Colloquium, Cambridge, England. (ERIC Document Reproduction Service ED 364 066).

Madaus, G. F. (1988). The influence of testing on curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education* (pp. 83-121). Chicago, IL: University of Chicago Press.

Magnan, S. (1982). Native speaker reaction as a criterion for error correction. ERIC Document Reproduction Service ED 22 066.

Maier, P. (1992) Politeness strategies in business letters by native and non-native English speakers, *English for Specific Purposes 11* (3):189-206.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Education Measurement: Issues and Practices, 8*, 14-22.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Mitchell, R., J. Hooper and F. Myles (in preparation) Prefabricated Phrases in Classroom Teaching and Learning: Their Place in Interlanguage Development. Presentation for AAAL 1997 Colloquium: Incorporating Sociolinguistic Perspectives into SLA Theory.

Mori, M.(1991) Script for the medical consultation: Comparison of the expectations between Japanese and American. Course paper, University of Minnesota.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

North, B. (1993). *The development of descriptors on scales of language proficiency.* Washington, DC: The National Foreign Language Center.

Oller, J. W. Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen, 7 6*, 165-174.

Olshtain, E. (in preparation) Pragmatic competence interacts with linguistic competence in various ways throughout the SLA continuum. Presentation for AAAL 1997 Colloquium: Incorporating Sociolinguistic Perspectives into SLA Theory.

Pavesi, M. (1987). Variability and systematicity in the acquisition of spatial prepositions. In R. Ellis (Ed.), *Second language acquisition in context* (pp. 67-72). Englewood Cliffs, NJ: Prentice-Hall International.

Piazza, L. G. (1980). French tolerance for grammatical errors made by Americans. *Modern Language Journal, 64*, 422-427.

Politzer, R. L. (178). Errors of English speakers of German as perceived and evaluated by German natives. *Modern Language Journal, 62*, 253-433.

Preston, D. (1989) *Sociolinguistics and SLA*. Blackwell.

Ranney, S. (1992) Learning a new script: An exploration of sociolinguistic competence. *Applied Linguistics* 13,1:25-49.

Rimarcik, J. (1996) AVRS telephone messages: Reality of the nightmare for international students. MA Qualifying Paper, ESL, University of Minnesota.

Sato, C. (1985). Task variation in interlanguage phonology. In S. Gass & C. Madden (Eds.), *Input in second language acquisition*, (pp. 181-196). Rowley, MA: Newbury House.

Savignon, S. J. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.

Schairer, K. E. (1992). Native speaker reaction to non-native speech. *Modern Language Journal, 76* (3), 309-319.

Schmidt, M. (1981) Needs assessment in English for Special Purposes: The case study, in L. Selinker, E. Tarone and V. Hanzeli (Eds.), *English for Academic and Technical Purposes: Studies in Honor of Louis Trimble*. Rowley, MA: Newbury House, pp. 199-210.

Schmidt, R. (1977). Sociolinguistic variation and language transfer in phonology. *Working Papers on Bilingualism, 12*, 79-95.

Selinker, L. and D. Douglas (1985) Wrestling with 'context' in interlanguage theory. *Applied Linguistics* 6:190-204.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedure. *Language Learning, 33* (4), 527-539.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*, 147-170.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal, 40* (3), 212-222.

Skehan, P. (1991). Progress in language testing: the 1990s. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 3-21). London: MacMillan Publishers Limited.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20* (5), 8-11.

Tarone, E. (1983) On the variability of interlanguage systems. *Applied Linguistics* 4,2:143-163.

Tarone, E. (1988) *Variation in Interlanguage.* London: Edward Arnold.

Tarone, E. (1990) On Variation in Interlanguage: A Response to Gregg," *Applied Linguistics 11*, 4:392-400.

Tarone, E. (1996a) Analyzing IL in natural settings: A sociolinguistic perspective on second-language acquisition. Paper presented at AILA, Jyvaskyla, Finland.

Tarone, E. (1996b) Variation and cognition: The impact of social factors on interlanguage construction. SLRF 1996 Colloquium: Integrating cognitive and linguistic approaches to SLA (Dean Mellow, organizer).

Tarone, E. and L. Beebe (in preparation) Incorporating a sociolinguistic perspective into SLA theory: Some questions we might address. Presentation for AAAL 1997 Colloquium: Incorporating Sociolinguistic Perspectives into SLA Theory.

Tarone, E. and G. Liu (1995) Should interlanguage variation be accounted for in a theory of second-language acquisition? In G. Cook and B. Seidlhofer (Eds), *Principles and Practice in Applied Linguistics: Studies in Honour of Henry Widdowson.* Oxford University Press.

Tarone, E. and G. Yule (1989). *Focus on the Language Learner.* Oxford: Oxford University press.

Thomas, M. (1994). Assessment of L2 proficiency in Second Language Acquisition Research. *Language Learning, 44,* 307-336.

Trites, J. (1996) English for sports purposes. M.A. Qualifying Paper, ESL, University of Minnesota, March 13, 1996.

Turner, C. & Upshur, J. (1995). Some effects of task type on the relation between communicative effectiveness and grammatical accuracy in intensive ESL classes. *TESL Canada Journal, 12* 18-31.

Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal, 49,* 3-12.

Wesche, M. (1992). Performance testing for work-related second language assessment. In E. Shohamy & R. Walton (Eds.), *Language Assessment for Feedback: Testing and Other Strategies* (pp. 103-122). Dubuque, IA: Kendall/Hunt Publishing Company.

Wolf, D. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal,77,* 473-489.

Young, R. (1991) *Variation in interlanguage morphology.* New York: Lang.

Young, R. (in preparation) SLA as discourse socialization. Presentation for AAAL 1997 Colloquium: Incorporating Sociolinguistic Perspectives into SLA Theory.